Media Corpora, Text Mining, and the Sociological Imagination - A Free Software Text Mining Approach to the Framing of Julian Assange by three news agencies using R.TeMiS

Gilles Bastin, Milan Bouchet-Valat

Media Corpora, Text Mining, and the Sociological Imagination

A Free Software Text Mining Approach to the Framing of Julian Assange by Three News Agencies Using R.TeMiS

Published in *Bulletin of Sociological Methodology*, 122, p. 5-25, 2014.

Preprint version. Please check the published version at
http://bms.sagepub.com/content/122/1/5.full.pdf+html

Gilles Bastin

Institute of Political Studies, Grenoble-Alpes University,
Pacte Research Centre (CNRS)
38041 Grenoble
France
+33 (0)4 76 82 61 04
gilles.bastin@iepg.fr

and

Milan Bouchet-Valat

Quantitative Sociology Laboratory (LSQ-CREST),
Center for Studies in Social Change (OSC-CNRS & Sciences Po Paris),
National Institute for Demographic Studies (INED)
France
nalimilan@club.fr

Media Corpora, Text Mining, and the Sociological Imagination

A Free Software Text Mining Approach to the Framing of Julian Assange by Three News Agencies Using R.TeMiS

Abstract

In this paper, we introduce R.TeMiS, a free software text mining solution aiming at exploring new dimensions in text mining with a particular focus on media framing analysis. R.TeMiS is especially designed to provide help in a) the automation of corpus construction and management procedures based on the use of large media content databases and b) the extension of the range of statistical tools available to social scientists exploring texts through R coding (one- and two-way tables, timelines, hierarchical clustering, content analysis, geographical mapping…). A case study on the media framing of Julian Assange from January 2010 to December 2011 is conducted. It is based on the analysis of a corpus of 667 news dispatches published in English by the three top international news agencies: Agence France Presse (AFP), Reuters and Associated Press (AP).

Keywords

Text mining, R, media studies, correspondence analysis, geographical mapping, framing, news agencies, Assange

Introduction

The methodology of applying statistics to large-n text corpora[1]—which we will here refer to as « text mining »—has gained legitimacy in every field of the humanities and social sciences (HSS) since the 1960s.[2] Many factors have contributed to this methodological shift. The rise in computer literacy among scholars, the diffusion of standard textual data analysis tools in surveys (Lebart et al., 1998) as well as the availability of software solutions based on graphical user interfaces providing such tools have notably diminished the costs of learning in this field. Simultaneously, significant changes have occurred in the way that social sciences access and store textual data. The digitalization of text corpora in many fields of social activity has for instance produced a deluge of available new documents to the social scientist, a phenomenon that has been referred to as a « data deluge » (Abbott, 2000; Hey and Trefethen, 2003).

Due to the digitalization of newspaper archives on the web, the availability of articles on platforms such as Lexis-Nexis or Factiva, the slow but growing textualization of audio-visual media, scholars working in media studies have particularly benefited from this new abundance of research materials and promising shift in methodologies.[3] Conventional content analysis — a method consisting of sampling newspaper archives and assigning abstract categories to passages of these articles in order to measure the recurrence of « themes » or « issues » within that content — has experienced a decline in this field of research.[4] It is increasingly challenged by new corpus-driven text mining techniques based on an exhaustive analysis of digitalized corpora extracted from databases, using keywords searches to select the material and inductive methods based on semantic (instead of thematic) fields identification within the corpus (frequency, concordances, collocation analysis).[5]

Unfortunately, the evolution of the software market has produced an abundant, intimidating and often obscure « maquis » of technical solutions for scholars intending to experiment with text mining methods on media corpora (Demazière and Brossaud, 2006). The average user must choose between dozens of applications. Moreover, each tool tends to be conceived as a black box: an integrated solution providing help from the corpus management to the edition of analysis reports, but also — and for that same reason — a world apart with its own peculiar customs, vocabulary and paradigms. As a consequence, text mining often appears as a very complex and clustered methodology, reserved only for those who share the theoretical paradigms behind the software they use.

This black box phenomenon obviously has some advantages when a researcher knows his box (the diminution of learning costs is one). But it clearly has also many disadvantages when it comes to the availability of such methods for researchers, the comparability of research results and the issue of controlling one's work in a way that helps, rather than hinders, the « sociological imagination » (Demazière, 2005; Mills, 1959). We strongly believe that using text mining as a means to test hypotheses and generate new interpretations of the social world—the « back-and-forth between sociological questioning and testing of

[1] By large-n text corpora we do not only mean text corpora containing a lot of words (like novels for instance) but also corpora made of a lot of different documents (like corpora made of hundreds of newspaper articles for instance).

2 In France for instance it has been acknowledged as a legitimate methodology in language science (Muller, 1969), history (Guerreau, 1989) and sociology. See for instance the preface of Baudelot to Lebart and Salem (1994).

3 As have scholars working with open-text responses to surveys (Bolden and Moscarola, 2000), scholars in the field of linguistics (Sinclair, 2004; Tognini-Bonelli, 2001) and discourse analysis (Baker, 2006).

4 See Berelson (1952) for the classical presentation of conventional content analysis and de Bonville (2000) for a recent survey. Interesting methodological discussions on conventional content analysis can be found in Krippendorff (2004a; 2004b) and Krippendorff and Bock (2008).

5 For an interesting discussion of the methodological issues of corpus-driven content analysis, see Schafraad (2006)

algorithms » (Demazière and Brossaud, 2006)—requires software solutions that can be fully controlled by users, rather than black boxes.[6]

The R statistical framework (R Core Team, 2013) has long been acknowledged to empower the social scientist with control over his or her work in every kind of statistical research. Thanks to the tm package (Feinerer, 2008; Feinerer et al., 2008; Feinerer, 2011), but also to other packages dedicated to advanced text mining operations as well as to general purpose packages which can be used for this particular application, R is a very fruitful environment for text mining. Yet, the power and the flexibility of this environment include the disadvantage that beginners can become confused due to the lack of a graphical user interface in the original R software (a feature that is especially interesting in text mining, in particular because text mining requires frequent close examination of the original corpus to contextualize statistical results).

In this article, we present R.TeMiS, a new R package that intends to fill this gap by providing users with a graphical interface (Bouchet-Valat and Bastin, 2013).[7] Even though R.TeMiS is a general purpose text mining tool, we will focus in this article on its use in media studies. The argumentation will thus be illustrated with a study on the media framing of Julian Assange from January 2010 to December 2011. This study is based on the analysis of a corpus of 667 news dispatches published in English by the three top international news agencies: Agence France Presse (AFP), Reuters and Associated Press (AP).[8] We do not aim here at a comprehensive analysis of this corpus but only to provide the reader with a good overview of R.TeMiS's possibilities using a real case study of media framing issues.[9] This case study has been chosen because the name of Julian Assange was associated at that period with competing « frames » in the news: the whistleblower who initiated the Wikileaks public scandals was one; the sexual offender who was facing a private scandal another.[10]

Moreover, analysis of dispatches offers a good challenge for text mining tools. Because news agencies provide news to other media outlets (newspapers, TV stations, radio, websites, etc.) they are essential

6 This is especially true in media studies. As Gerbner once put it, in this field of research, social scientists aim at finding « hidden regularities of content which record and reflect objective mechanisms of a social order ». Therefore, they need complete and reflexive control over their methodological tools, meaning the ability to easily gather and process big corpora, apply robust statistical methodology to those corpora, and to manipulate contextual variables for understanding the contents and easily use the data for a range of new treatments. « The classical role of cultural scholarship as a testing ground of critical social theory, Gerbner concluded, is to be strengthened, broadened, and deepened — not abolished — in the analysis of mass media content through the newer, more systematic and refined methodologies » (Gerbner, 1958).

7 R.TeMiS has been developed as a specific menu of the R Commander, a well-known and very robust GUI for R (Fox, 2005). For that reason, the package name in R conforms to the Rcmdr plugins syntax («RcmdrPlugin.temis») but can be shortened as R.TeMiS (for R Text Mining Solution). R.TeMiS also refers to Artemis, the sister of Apollo, a goddess who was highly venerated in ancient Greece. She was often depicted as a huntress wandering in forests, uncultivated areas and other wild lands. Instructions to download and launch R.TeMiS are available at the following URL: http://rtemis.hypotheses.org/. For a general overview of the package's features, see Bouchet-Valat and Bastin (2013).

8 The search was made on Factiva, a Dow Jones company providing access to media contents. Every dispatch containing the term "Assange" in the headline or 1st paragraph was retained.

9 News «framing» has been widely studied since the 1980's. The paradigm has many possible interpretations and has even been called a « fractured paradigm » (Entman, 1993). We will use « frames » here in a very minimal and commonly shared sense as « the central organizing idea or story line that provides meaning to an unfolding series of events » (Gamson and Modigliani, 1987: 143). For a further review of the framing literature, see De Vreese (2005).

10 Julian Assange co-founded Wikileaks in 2006. Since then, the website has specialized in leaking secret files such as the famous Afghan and Iraqi « war logs » and secret US cables in 2010. Julian Assange has received various prizes for his contribution to freedom of the press and was elected Person of the Year by Time magazine and Le Monde in 2010. The same year, the site began to be targeted by the American government, which succeeded in blocking its financial resources (Visa and Paypal). On the 20th of August 2010, Julian Assange was accused of sexual offenses in Sweden. On the 8th of December 2010 he was arrested in Great-Britain, then was granted bail after the payment of 240.000 £ in cash and sureties. In June 2012 Assange violated the conditions of his bail and applied for political asylum at the Ecuadorean Embassy in London.

actors in the early framing of social and political problems. At the same time, since they do it in very routinized forms—in very « objective » forms to use the journalistic vocabulary—their coverage of breaking news stories like the one we study is very similar. Thus, identifying the frames they produce requires precise scrutiny of the vocabulary employed in their dispatches. Briefly, elucidating framing issues in news agencies' coverage of topics like Julian Assange is a good way to test the ability of text mining tools to identify small but meaningful differences in large-n corpora that are easily constituted from existing electronic sources. This facilitates discovery of what contemporary media studies can expect from text mining.

This paper is composed of six sections. In the first section, we highlight three reasons to choose free software in the field of media corpora text mining. In the second section, we introduce the importation, coding and management of media text corpora in R.TeMiS. In the third section, we illustrate the package's variables visualization features. The fourth section is devoted to elementary statistics with R.TeMiS. The fifth section addresses hierarchical clustering and correspondence analysis. The sixth section provides an illustration of the advantages of R for extending the range of statistical procedures that are currently available (with a focus on geographical mapping of term frequencies).

Advantages of a free software text mining approach to media studies

Opting for free software or open-source statistical solutions has not only to do with broad preferences for costless software solutions. It has also to do with workflow choices that have an effect on the results produced. Three main reasons can be identified for preferring an open-source approach: free cost, robustness and reusability.

The first advantage of open-source text mining is free cost. Most available text mining software solutions are sold at a high price point, whether as stand-alone processing solutions or as add-ons to general purpose—and expansive—statistical software solutions.[11] This is clearly a limitation for their use, and obviously presents difficulties for researchers without access to financial resources (not to mention graduate and PhD students or researchers in emerging countries). Free cost also presents another advantage. Whereas proprietary software solutions tend to be arbitrarily limited in their processing capacities (according to the kind of license paid), free software solutions do not represent such limits outside the user's hardware processing capacity and time available (of course very big corpora can induce a very long calculation time). Proprietary text mining software solutions typically tend to limit the maximal size of corpora, sometimes with different limits according to the license paid. This is a big issue in media research where corpora can be significant in size. This is evident in situations where researchers adapt the size of their corpus to fit the software's own limitations, rather than research questions.

Another interesting feature of open-source text mining is robustness. Most packages in R have been created by very experienced researchers in their field. They are constantly scrutinized and improved by users' communities. In contrast, closed source text mining software solutions do not benefit from this kind of community-driven improvements. Each proprietary text mining application must include its own implementation of fundamental statistical methods. This results in a risk of introducing bugs that can go unnoticed since the code cannot be checked and the same analysis cannot always be performed using different programs to compare results.

11 Some are distributed for free, especially those developed by small teams of researchers. But such small dedicated tools are usually closed-source and have difficulty competing with commercial solutions since they need to re-implement every feature from scratch despite limited resources.

The third key feature of free software text mining is reusability. Integration with a free general-purpose statistical framework such as R means that all the methods developed for this environment are made available to text mining. By default, R.TeMiS makes use of many free software packages[12], which have in some cases been modified by their authors to better suit their needs. Even more features are offered to advanced users by giving them access via manual code edition to any method they want. This is congruent with the spirit of free software, and contrary to proprietary software solutions that do not allow users to modify software code to fit specific scientific needs (outside the usual parameters toolboxes). R.TeMiS thus gives the user full control over the code producing the analyses. Following the general principle adopted by the R Commander, the commands generated by the user's actions in dialog boxes are printed in a script box and can be edited and run as plain code. This allows users to check what is being done by the software and optionally extend standard analyses with custom R commands most fitted to their needs.

Importing, coding and managing large-n media corpora with R.TeMiS

An important feature of the R.TeMiS approach to media studies is to facilitate the constitution of media corpora. Handling and coding corpora is a key moment in media research. Yet, too often, it is considered only as a preliminary step, and not part of the analysis. Text mining solutions typically require the researcher to produce specific types of documents (or files) with a specific formatting that can only be obtained manually. Consequently, researchers have to perform operations such as copy/paste chunks of text in order to transform many documents into one (e.g., many newspaper articles pasted in one single document in a chronological order), insert code lines within the corpus to introduce contextual variables (e.g., date of publication, source, author), suppress certain undesired recurring words, or expressions that hinder the analysis (such as the title of the source media).

The disadvantages of such handmade corpora are evident. The first is the vast amount of time needed to constitute the corpus. Time considerations may compel researchers to outsource this analytical step and thus, at least partially, surrender control. A second disadvantage is the high probability of error during manual manipulation of the corpus (e.g., deletion of documents, errors in automatic replacement formulas, etc.) or losing track of the changes performed (because users modify their source corpus, changes cannot be undone unless a very precise record has been kept). Last but not least, such corpora produce locked-in effects due to dependency on the software used.[13]

The integrated approach that has been adopted in R.TeMiS consists of providing the user with corpus importation filters that allow the use of many kind of corpora without customized intervention by the researcher. These filters simplify the characterization of the contents with relevant metadata (date, author, source, etc.) by automatically retrieving variables when they are available (as is often the case in structured media contents databases) or by allowing the researcher to document his corpus without modifying it. Whenever necessary the R.TeMiS package provides tools for traceable corpus changes from within the software (i.e. without manually modifying the corpus).

To do so, R.TeMiS supports different kinds of structured source files corresponding to various kinds of research materials:

12This is notably the case for stemming and, for corpora importation, but also for graphics, which are usually very poor in proprietary solutions but of very high quality in R.

13 For example, when variables are added manually within the corpus, using a syntax that is specific to the software (thus making it difficult to perform different kind of analysis without recoding the corpus and then facing the waste of time and high probability of errors issues).

• series of plain text files (.txt) contained in a directory (typically the result of audiovisual media materials transcriptions or a sociological interviews campaign);

• spreadsheet-like files (.csv, .ods, .xls) with one line per document/person. The first column containing the text to analyze and the remaining columns providing information about the document (typically the answers to a free-form question in a survey research or tabulated lists of newspaper articles containing columns for text and relevant variables);

• structured files (.xml or .html) exported from the Dow Jones Factiva content provider (typically the result of a media content analysis process using keyword search);

• Twitter searches on hashtags, authors or full text, or more complex queries using the Twitter API.[14]

The Assange corpus has for instance been downloaded from Factiva as a series of seven structured .html files containing dispatches by bunches of one hundred. All files have been saved in one directory on the computer. The R.TeMiS 'Import corpus' dialog box (Fig. 1) is used to process the documents (the dispatches). This action produces the document-term matrix (DTM) that is the basis of all further statistical treatments. Documents can be split into smaller ones that are defined as a number of adjacent paragraphs (every dispatch would then be separated into many documents). This feature can be useful for relatively long and potentially heterogeneous documents.[15] Documents can also be processed to make them more suitable for statistical analysis : conversion of the text to lower case, punctuation removal, stopwords removal, and finally stemming that is carried out using language-specific algorithms derived from the work of Martin Porter.[16.] This processing option is most interesting for small corpora since slight grammatical variations can then reduce the number of co-occurrences between two documents. While importing the corpus, R.TeMiS automatically retrieves variables from the Factiva .html files and asks the user which should be retained. Only two are kept in the following: the Origin (AFP, AP or Reuters) and the Date (in a YY-MM-DD format).

14 Many other corpus importation filters can be developed pertaining to every user's need. R.TeMiS automatically retrieves variables in the case of spreadsheet files, structured Factiva files and Twitter searches. For plain text files, the user only has to fill variables information for each document or provide the R Commander with a spreadsheet-like file with one row per document (identified by its order) and one column per variable.

15 Corpus heterogeneity can stem from complex factors: in some cases keyword searches can produce artifacts when the same term refers to very different things; media contents can also be indexed strangely in databases, such as grouping all « News in Brief » kinds of articles together, regardless of subject. Splitting documents in smaller ones based on paragraphs and choosing only the paragraphs containing specific terms can help to solve this problem.

16 These algorithms are provided by the Snowball project, cf. http://snowball.tartarus.org/, and used in R via the SnowballC package (Bouchet-Valat, 2013). Stemming is currently supported for the following languages: Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Norwegian, Portuguese, Romanian, Russian, Spanish, Swedish and Turkish.

Fig. 1 — The 'Import corpus...' toolbox while importing the Assange corpus

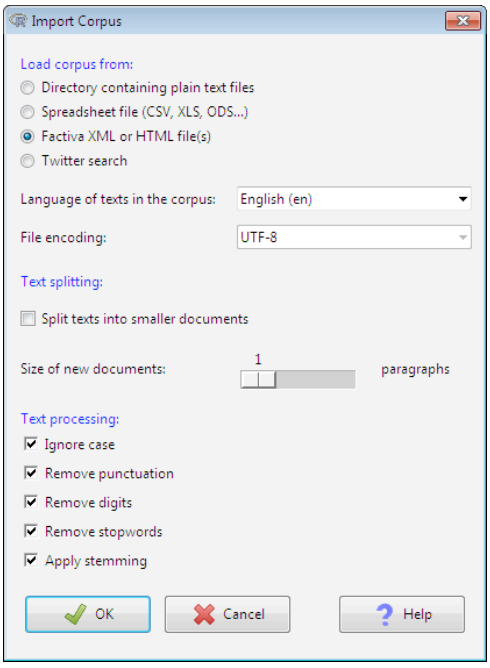
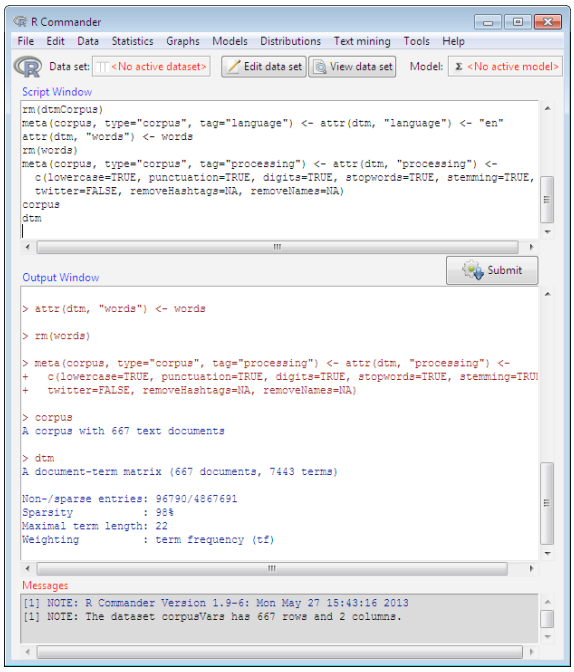


Fig. 2 — Results of the corpus importation



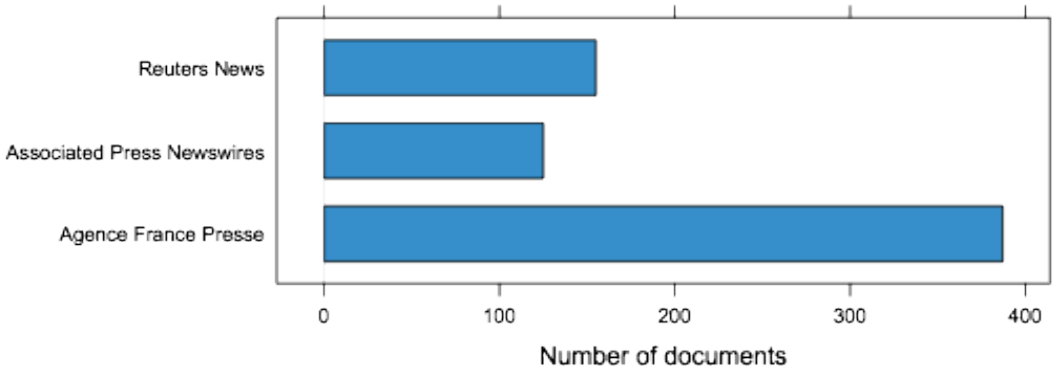
After this step, a summary of the DTM is printed on the screen (see Fig. 2). The number of documents (lines of the DTM) and the number of terms (columns of the DTM) are printed, as well as a measure of the matrix sparsity (% of cells with zero occurrences) and the weighting unit. Later analyses are performed using the 'Text mining' menu in the Rcmdr window.

Visualizing relationships among variables within the corpus: the time and source structure behind the coverage of the Assange case

Media analysis often requires the manipulation of contextual variables describing the documents before launching textual data exploration within documents. For that purpose, R.TeMiS provides users with meta-data visualization tools such as one- and two-way tables using meta-data variables (with optional plotting of the results). Plotting the number of dispatches produced by each news agency is interesting for the Assange case (Fig. 3). This plot clearly shows that AFP devoted far more coverage than Reuters and AP to Julian Assange in 2010-2011. Whether this is due to editorial choices putting Assange on the media

agenda more often or artifactual factors linked to different dispatches publication policies is of course impossible to say without further qualitative analysis of the corpus. However, as far as the raw number of dispatches is concerned, AFP clearly gave more visibility to Julian Assange than did its competitors.

Fig. 3 — Coverage of the Assange case by the three agencies



The same menu also makes it possible to plot time series representing the number of documents over time, using a single curve, or one curve for each level of a variable. A rolling mean can be computed over a configurable time window.[17] The Assange corpus already contains date information imported from Factiva, which is very useful in studying media cycles around this subject. For example, the two following figures clearly show a concentration of media attention in December 2010, when Assange was arrested in Great Britain, and not in August when the sexual assault case began in Sweden (Fig. 4 shows daily counts whereas Fig. 5 includes a rolling mean to attenuate the very important daily variations of such kind of media contents). These Figures clearly show the « episodic » nature of the media framing of Julian Assange.[18]

17 This feature is based on two state-of-the-art R packages: zoo for time series handling (Zeileis and Grothendieck, 2005) and lattice for plotting (Sarkar, 2008). This again means that the generated code can easily be extended for custom representations if needed.

18 According to Shanto Iyengar, episodic frames contrast with thematic frames because they tend to focus the attention on one event (or person) and divert it from a broader context and the public issues underlying the event or person (Iyengar, 1994).

Fig. 4 — A timeline of all documents in the Assange corpus by source

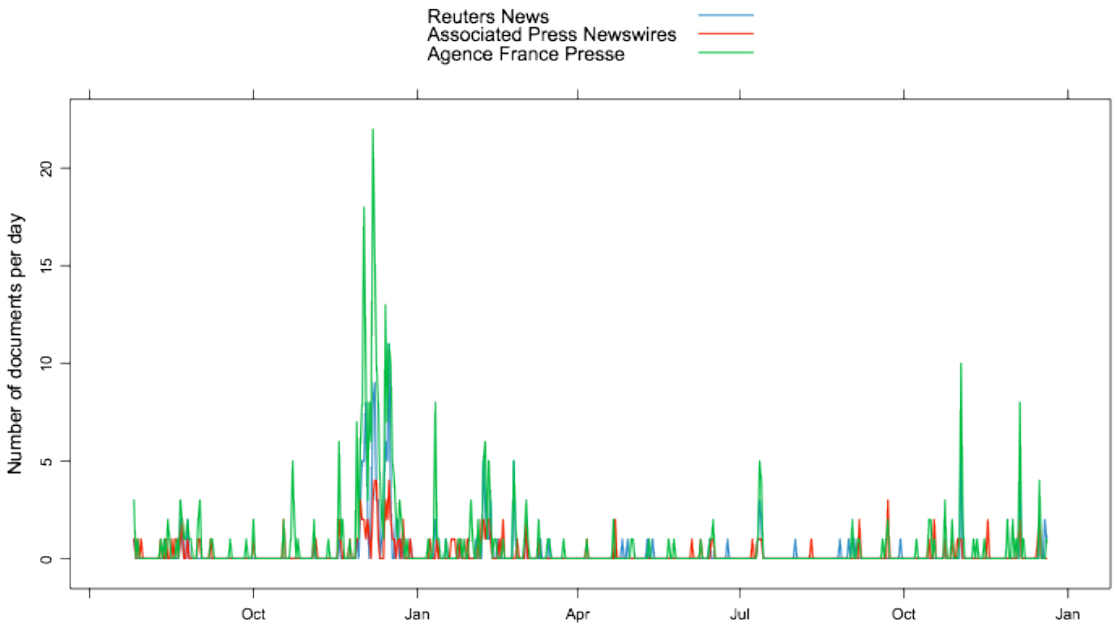
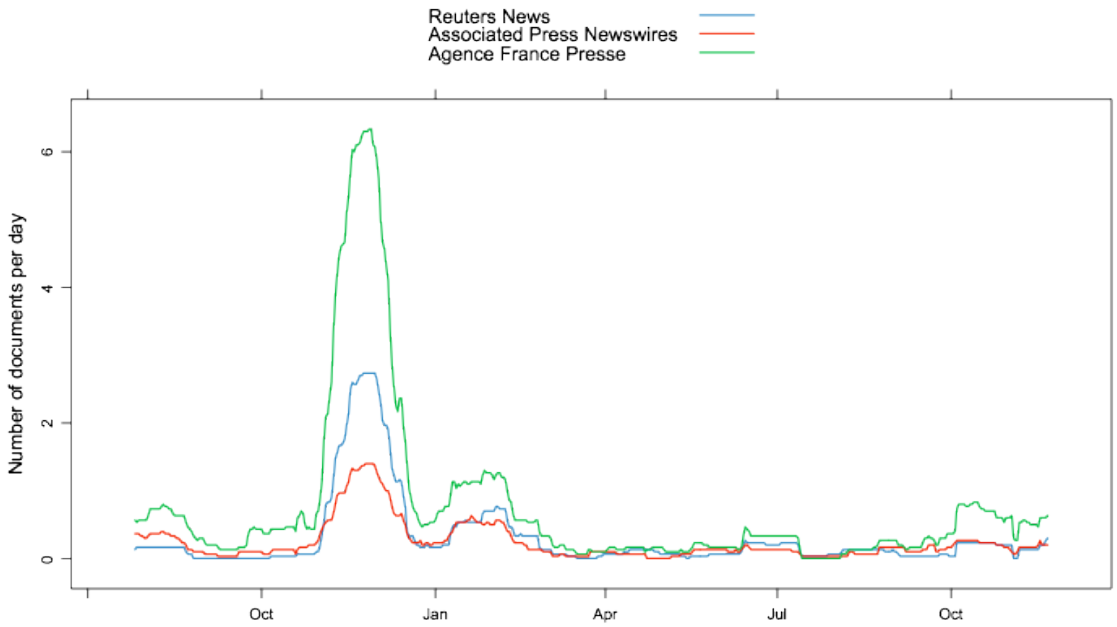


Fig. 5 — A timeline of all documents in the Assange corpus by source with a 30 day rolling mean



A more conventional two-way table representation of the media chronology using a recoding of the Date variable to monthly breaks instead of daily ones is also very useful to assess the level of episodicism of every news agency.[19] As can be observed in Table 1 below, Reuters demonstrated the most concentrated time structure among the three agencies. The coverage for December 2010 (Julian Assange's surrender to the London police and granting of bail by the High Court) and February 2011 (a District Judge in south London ruled that Assange should be extradited to Sweden) accounted for 65% of the overall coverage (54% for AFP and 42% for AP). AP seemed to provide the most consistent coverage.

Table 1 — News agency distribution of dispatches on Julian Assange in 2010-2011 (monthly breaks)

	Agence France Presse	Associated Press Newswires	Reuters News
2010-07	1.03	2.40	0.65
2010-08	4.91	8.00	3.23
2010-09	1.81	1.60	0.00
2010-10	3.36	2.40	0.00
2010-11	6.98	8.80	3.87
2010-12	45.48	30.40	50.97

19 The recoding of the Date variable is performed with the menu "Manage corpus → Recode time variable". The table itself is available in the "Distribution of documents → Two-way table of variables" menu.

2011-01	5.17	5.60	3.23
2011-02	8.79	12.00	13.55
2011-03	2.58	2.40	2.58
2011-04	0.78	2.40	1.29
2011-05	1.29	0.00	1.29
2011-06	1.29	3.20	1.29
2011-07	2.58	3.20	3.87
2011-08	0.00	0.80	1.94
2011-09	2.07	4.00	1.94
2011-10	2.84	4.80	0.65
2011-11	4.65	3.20	3.87
2011-12	4.39	4.80	5.81
Sum	100.00	100.00	100.00

Identifying media frames through elementary corpus statistics

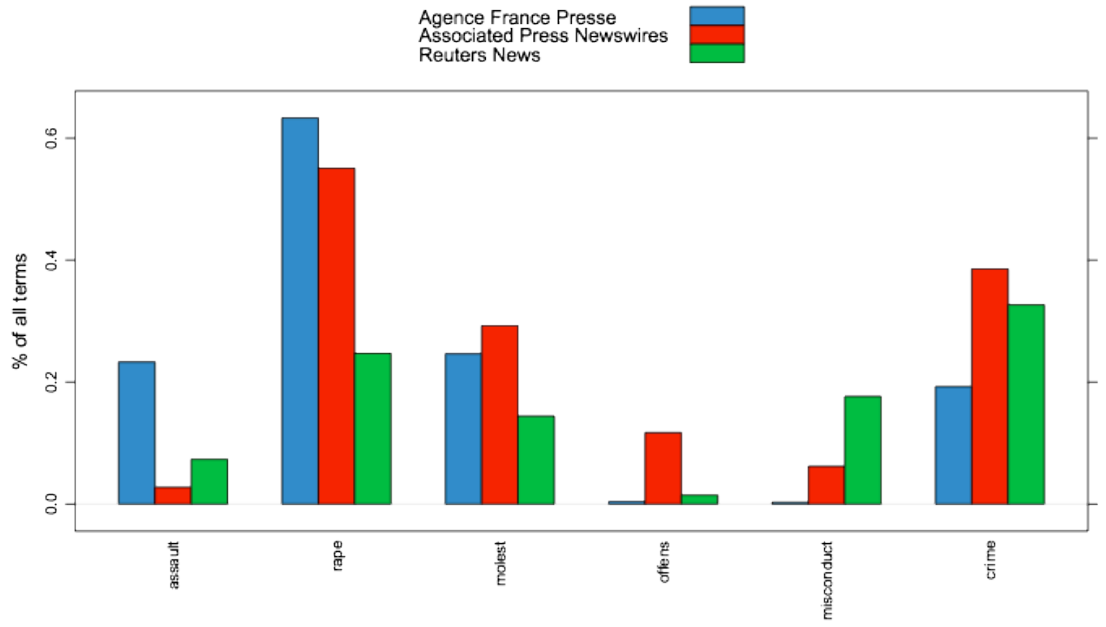
Media frames can be identified in a corpus by finding those terms that are very specific of a given level of a contextual variable – i.e. terms whose observed frequency in each level is either too high or too low compared to what would be expected given the documents' lengths and the global distribution of terms in the corpus. If we consider for instance the three different sources gathered in the Assange corpus, some meaningful differences arise from specific terms analysis (see Appendix 1).[20] The use of « whistleblower » for instance is very specific to AFP (referring to Assange's role as leader of Wikileaks) and is specifically absent from the two other sources. The different terms used to describe the alleged rape (« assault » and « rape ») were specific to AFP in contrast with « offense » (specific to AP) and « misconduct » or « crime » (specific to Reuters). Some secondary stories within the Assange story also seemed to have stronger connections with a specific other agency : AP for instance showed significant use of « Mayawati », the name of Uttar Pradesh's chief minister who was involved in the Cablegate leak. Reuters significantly mentioned « Elmer », the name of a former Swiss banker who collaborated with Wikileaks and even gave a press conference with Assange in January 2011. This interestingly suggests that press agencies have some autonomy in choosing angles and framing what happened when compared to the « primary definition » (Hall, 1978) performed by the Swedish prosecutor Marianne Ny. The fact that her decision was written in Swedish of course opened up space for interpretation, notably of the Swedish term « sexuellt ofredande » that can be translated many ways.

Instead of focusing only on statistically specific terms, researchers working in media studies can preferentially examine only those terms that are relevant to their research question and hypothesis. Indeed, such terms can be sociological artifacts (like the very specific occurrence of press agencies acronyms in their dispatches), and with large corpora there are so many significantly specific terms that the researcher can-

20 All terms relating to the agencies' names, reporters' names or bylines, edition process and dates have been removed from the corpus using the 'Manage corpus → Select or exclude terms' menu (the list is the following: "id," "ad," "dk," "gj," "ar," "satter," "raphael," "nn," "edit," "afp," "novemb," "decemb," "rdm," "sr," "ap," "associ," "contribut," "http," "dec," "nov," "aug," "reuter," "feb," "rjm," "nl," "holden," "davi," "shanley," "hosenbal," "bur," "mark," "david," "michael," "adrian," "keith," "mia," "patrick," "stefano," "ambrogi," "ga," "mb," "writer," "press"). Specific terms analysis ('Descriptive analysis of vocabulary → Terms specific of levels... ') relies on the measure of p-values based on an hypergeometric distribution that give the probability of observing such extreme number of occurrences in the level under the independence hypothesis; the sign of t-values can be used to identify positive (printed first) and negative (printed last) associations.

not analye all of them. R.TeMiS offers interesting features for a deductive approach based on analysing terms chosen by the researcher as well. The previous measures can be computed for chosen terms by levels of a variable and a plot can be drawn (see Appendix 2 and Fig. 6 for an application of terms used in the accusations against Assange). For instance, the terms « molestation », « molested », « molesting » — stemmed as « molest » — weren't specific enough to appear among the 25 most specific terms in the previous treatment. They can still be considered interesting because they are closest in meaning to the terms used by the Swedish prosecutor.[21] This analysis strongly contrasts with AFP and Reuters which did not use a similar vocabulary to describe the charges.

Fig. 6 — Framing the facts in the Assange corpus



An inductive approach to media framing using correspondence analysis

Hierarchical Clustering (HC) and Correspondence Analysis (CA) are two very popular ways of handling text corpora and trying to reveal their structure directly from the kind of vocabulary employed. But performing such analysis with big media corpora that has been automatically retrieved from databases or web archives induces some methodological peculiarities due to various sources of heterogeneity and redundancy within the corpus. R.TeMiS provides many computing and visualizing parameters that have been designed to help the user create the best possible representation of his corpus (not taking into ac-

21 Other measures can be computed by R.TeMiS, both by document and by levels of a variable: vocabulary summary (number of terms, diversity and complexity) ; co-occurring terms (terms that tend to appear in the same documents as a chosen term); dissimilarity table (Chi-squared distance between row profiles of the document-term matrix collapsed according to the levels of a variable).

count the unlimited possibilities offered by a direct access to the R code). In what follows CA will be used to illustrate an inductive way to identify frames within the Assange corpus.[22]

A first issue in large-n Factiva corpora is the artifacts created by the presence of irrelevant terms in parts of the corpus. In the Assange corpus it appears that the recurrence of the term « ID » in Reuters documents (a reference to other dispatches identification number within the database) produces artifactual results that are very obvious in the first CA produced if one does not intervene on the corpus. Previously excluding the term is of course the best thing to do. When dealing with large corpora, it can also be very useful to limit the number of terms taken into account and thus dramatically reduce the computing requirements of CA and HC in terms of both memory and time. R.TeMiS thus offers an option to exclude terms that are not present in more than a given percentage of the documents (sparsity level). For instance, using a 96% sparsity parameter leads to omitting terms absent from more than 4% of the documents — meaning 27 in the Assange corpus—in the DTM. The user can also limit the number of terms plotted on a CA graph to the most contributing ones.[23]

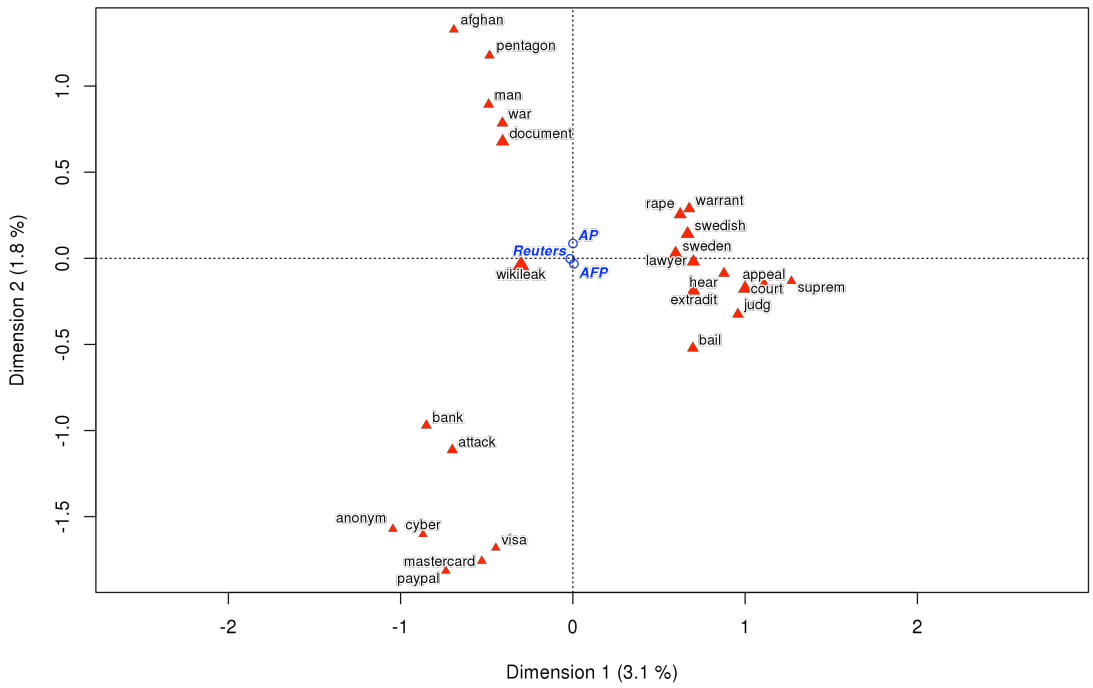
Fig. 7 shows the result of a CA conducted on the Assange corpus (with deletion of all terms mentioned in note 18), a 96% sparsity parameter and selection of the 30 most contributing terms to both axes only. The first plane of this correspondence analysis, despite the low level of total variance explained (which is due to the large amount of information contained in a large corpus), highlights the presence of three main media frames influencing Julian Assange's media coverage during the studied period: the Swedish « rape » case and (essentially) its British and Swedish judiciary consequences (the ten most contributive terms on the positive side of axis 1 are « court », « appeal », « swedish », « extradit », « sweden », « lawyer », « judg », « hear », « rape », « suprem ») ; the Wikileaks context is behind the most contributive terms in the top left quadrant (« document », « war », « pentagon », « afghan », « man », « civilian », « iraqi », « militari », « classifi », « afghanistan ») ; the financial surroundings of the Wikileaks issue and the attempts of the American administration to block Wikileaks' financial resources is suggested by the most contributive terms in the bottom left quadrant (« attack », « mastercard », « visa », « bank », « paypal », « anonym », « cyber », « payment », « swiss », « compani »). [24] This third frame would have been very hard to identify in the corpus without fine tuning of the sparsity permitted by R.TeMiS.

Fig. 7 — The three frames identified by CA on the Assange corpus

22 The R.TeMiS menu provides tools to perform both CA and HC.

23 R.TeMiS users can also decide to restrict the corpus to a specific subset of documents, by retaining only those containing (or not containing) some terms, or those corresponding to the given level of a variable. Because « documents » can either mean the initial text documents (here the dispatches) or text chunks within these documents (if this option was checked during the import process), thematic corpora can be created from large and heterogeneous original corpora: a selection of chunks (for instance paragraphs) according to a set of terms contained in these paragraphs can help focus the corpus.

24 While showing the results of a CA, R.TeMiS opens a new window containing summary information on the plotted axes: percentage of inertia, most contributive terms on the positive and negative side, most contributive documents on each side (with the text contained in these documents), and position and quality of representation of the variables on the axis. This is very useful to interpret the CA planes.



News agencies do not really differenciate on this first plane. When plotted as levels of a supplementary variable, they are located right in the middle of the first plane. A way of trying to find differences between agencies is to produce a new CA based on a version of the DTM aggregated by the levels of the Origin variable, a feature that is proposed in R.TeMiS's CA menu. This classical method allows imposition of an interpretation framework defined a priori, making more apparent the differences that are of interest for the question at hand.

The first axis of this new CA clearly opposes AFP on the negative side to Reuters and AP on the positive side (see Fig. 8 below). AFP contributes to 45% of the construction of this axis and has a 100% representation quality on it. AP contributes to 30% of the axis construction and has a 47% representation quality. Reuters contributes to 25% and has a 42% representation quality.[25]

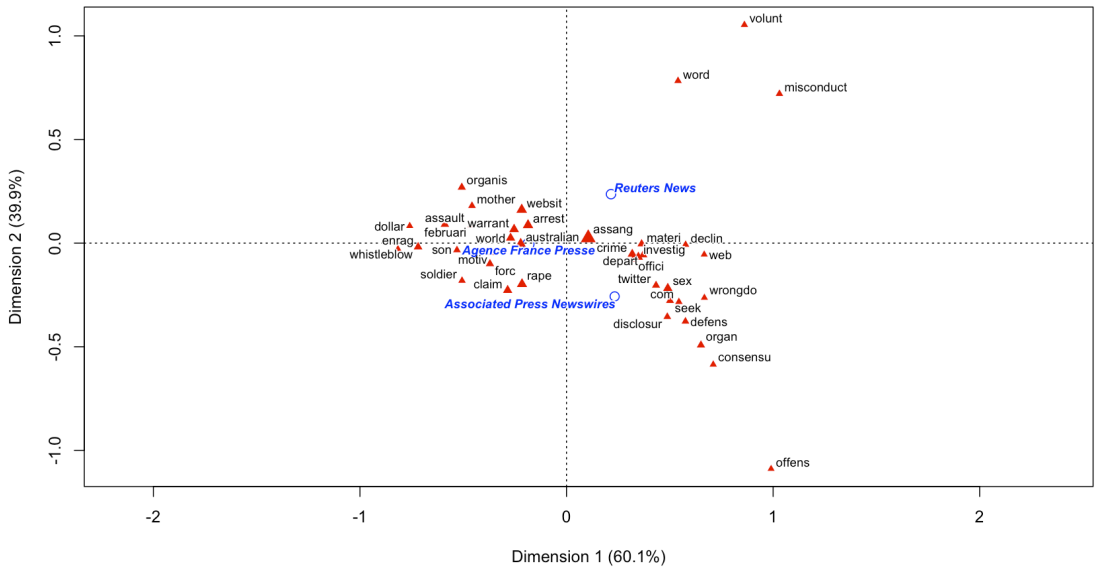
Terms that stand on the negative side of this axis with AFP are « whistleblow » (3.09%), « assault » (1.68), « organis » (1.04), « websit » (0.97), « warrant » (0.82), « rape » (0.78), « claim » (0.76), « australian » (0.76), « dollar » (0.72) and « enrag » (0.69). In addition to translation choices concerning the facts (« rape » and « assault ») and the already mentioned use of « whistleblower » to describe Assange and his « website » and « organisation », these terms express the pressure directed at Assanges by the judiciary process (« warrant » but also « arrest » (0.66)) and a personal account of his situation with the use of references to his being Australian. His family (« son » 0.63) and « mother » (0.48) appear on this side of the axis due to dispatches mentioning statements made by his mother concerning various aspects

25With only three news agencies, only two CA axes are by definition required to describe 100% of the variance.

of the case. « dollar » has been used by Assange himself to blame the US for the losses in financial support to Wikileaks following the funding blockade. AFP also published numerous dispatches mentioning that Assange had « enraged » US authorities by releasing classified documents.

AP and Reuters are positionned on the positive side of axis 1 with the following contributive terms : « sex » (2.33), « misconduct » (2.00), « organ » (1.76), « assang » (1.23), « investig » (1.17), « volunt » (1.17), « offens » (0.97), « defens » (0.79), « offici » (0.78). The two agencies are connected to a different frame concerning the facts. The terms chosen to describe them differ with an emphasis on the sexual nature of those facts and a more moralistic perspective (« consensual », « misconduct », « offense » but also « wrongdoing » (0.52)). « Volunt » stands for « volunteers » and describes the status of the two women inside the Wikileaks organization. The legal procedure is also highlighted with references to « investigations » and Assange's « defense ». By comparison with the personal tone used on the negative side of this axis, here « officials » (and « official » statements) are more present.

Fig. 8 — CA on DTM aggregated by Origin (terms most contributive to axis 1)

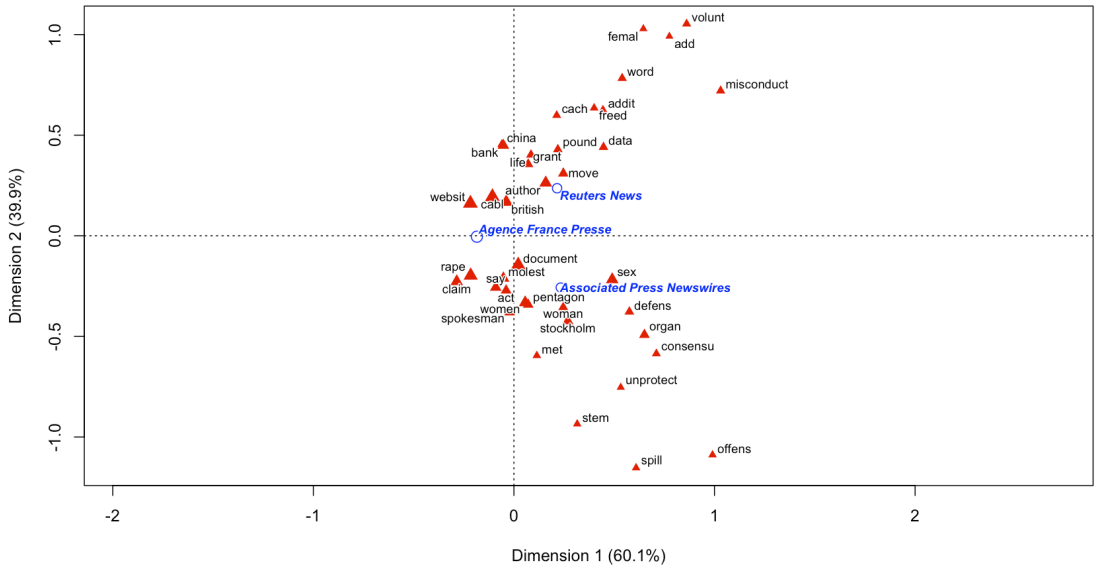


The second axis opposes Reuters and AP (see Fig. 9 below). Reuters contributes to 51% of the axis construction and has a 58% representation quality on this axis. It stands on the positive side with terms like « volunt » (2.63%), « bank » (1.95), « word » (1.78), « misconduct » (1.47), « femal » (1.24), « cabl » (1.20), « author » (1.06), « add » (0.94), « websit » (0.81), « cach » (0.76)[26]. AP contributes to 49% of this axis and has a 53% representation quality. It stands on the negative side with significant terms like « spill » (1.98)[27], « offens » (1.76), « women » (1.55), « organ » (1.52)[28], « stockholm » (1.36), « stem »

26 Referring to documents or cable « caches ».
27 Referring to Wikileaks as a « secret-spilling website ».

(1.30), « rape » (0.98), « spokesman » (0.83), « claim » (0.73). In addition to a clear focus on Sweden, this side of the axis is more concerned with the sexual case and expresses a formal use of objectivity « rituals » (Tuchman, 1972) like mentioning sources and using quotes (« spokesman », « claim », « document » (0.67), « say » (0.61). « Pentagon » also appears there, which is congruent with AP's national origin.

Fig. 9 — CA on DTM aggregated by Origin (terms most contributive to axis 2)



Mapping Julian Assange?

Due to the fact that R.TeMiS edits plain R code at every step of the lexical analysis, it enables using this code to extend the range of statistical procedures applied to the document-text matrix beyond classical text mining. Every kind of statistics or visualization method can be applied to the DTM. The Assange corpus offers a good example of this feature. The framing of Assange by the three news agencies entails a geographical dimension: media frames, in this case, are also about drawing a world map to portray Assange. The rape case for instance mostly connects Assange to Sweden and Great Britain. Other frames

28 « organ » and « organis » are root stems for terms like « organize » or « organization ». The only difference comes from the fact that they can be written with s (then stemmed as « organis ») or z (then stemmed as « organ »), i.e. using respectively the English or American spellings.

such as the Wikileaks context connect Assange to other parts of the world: Australia where Assange was born, Iraq and Afghanistan referring to Wikileaks' massive leaks of classified information in 2010, the United States due to its determination to stop Wikileaks from operating, etc. Plotting a world map of the Assange case according to the three agencies provides an interesting way to visualize this corpus.[29]

Frequencies of terms referring to a country can of course be tabulated within R.TeMiS. After identifying those terms using the 'Terms dictionary' menu (a menu that is also useful to check how words have been stemmed), their frequencies can be computed using the 'Descriptive analysis of vocabulary → Analysis of chosen terms...' menu. Appendix 3 contains the results of this analysis for each level of the Origin variable.[30] Some interesting things appear clearly in this table such as the importance of references to Sweden (« sweden », « swedish », « stockholm ») and Great Britain (« britain », « british », « london », « england », « english », « uk ») that exceed every other geographical denomination and exceed 0.3% of the occurrences at least one news agency. The only other term that shows the same level of frequency is « australian ». The t-values can be used in these tables to identify significant over- and underrepresented terms in the three agencies. The three terms referring to Sweden are for instance under-represented in AFP (that strongly contrasts with AP on this point). To the contrary, AFP makes significant use of terms referring to Australia—hence to Julian Assange's biography—in comparison to both AP and Reuters.

These tables have strong limitations. The first involves the various terms used to refer to the same country (country names, city names and adjectives). The second involves the fact that comparisons are hindered by the production of three different tables instead of one. A little piece of R code, written directly in the Rcmdr script window, can help by constructing groups of terms and calculating frequencies for these groups. In Table 2 for instance, terms have been grouped using two-digit country names as level.[31]

Table 2 — The geography of the Assange corpus (frequencies by country)

	Agence France Presse	Associated Press Newswires	Reuters News
SE	1.6801	2.1796	1.7514
GB	1.4403	1.3084	1.6043
AF	0.3193	0.3443	0.2149
IQ	0.2735	0.2135	0.2237
AU	0.7087	0.4511	0.4474
US	0.4621	0.5647	0.5505
IS	0.0593	0.0448	0.0353

29 Thinking that way when studying international news agencies is also interesting because news agencies are often suspected of being biased according to their national origin.

30 Every geographical term with more than 20 occurrences in its stemmed form in the whole corpus has been retained. Terms are ordered by country.

31 The code is the following :

```
getPercents <- function(terms) colSums(rbind(termFrequencies(dtm, terms, meta(corpus, "Origin")[[1]])[, 1,]))
countries <- list(SE=c("sweden", "swedish", "stockholm"), UK=c("britain", "british", "london", "england", "english", "uk"), AF=c("afghanistan", "afghan"), IQ=c("iraq", "iraqi", "baghdad"), AU= c("australia", "australian"), US=c("america", "american", "washington", "york", "usa"), IC=c("iceland"), RU=c("russia", "russian"), CH=c("switzerland", "swiss"), DE=c("germani", "german", "berlin"), FR=c("franc", "french"), IR=c("iran"), PK=c("pakistan"), EC=c("ecuador"), ES=c("spain"))
freqs <- t(sapply(countries, getPercents))
freqs
```

RU	0.0808	0.0482	0.0736
CH	0.1415	0.1412	0.1531
DE	0.0714	0.1653	0.1030
FR	0.0768	0.0413	0.0353
IR	0.0162	0.0207	0.0589
PK	0.0310	0.0069	0.0088
EC	0.0216	0.0034	0.0206
ES	0.0202	0.0103	0.0059

Small differences can be observed in Table 2 in the way that news agencies mapped the Assange case. When compared to its European counterpart Reuters, AFP highlighted countries such as Afghanistan, Iraq, Australia, Iceland, Russia, France, Pakistan, Ecuador and Spain. But it relatively under-reported for Sweden, Great Britain, the US and Germany. AP over-reported the U.S. as could be expected but under-reported Great Britain (on the contrary, Reuters over-reported both the US and the UK). A little R coding can help to geographically plot these data. R provides packages, like the maps package (Becker et al., 2013), that can easily draw maps and plot points using standard geographical coordinates. Transforming Table 2 into a data frame and adding two new variables (a latitude and a longitude for each country) is all we require. It can be done manually but the risk of making errors would then be significant. Instead, finding a data frame with country centroïds coordinates and importing it in R to merge it with the frequencies data frame is a better option.[32]

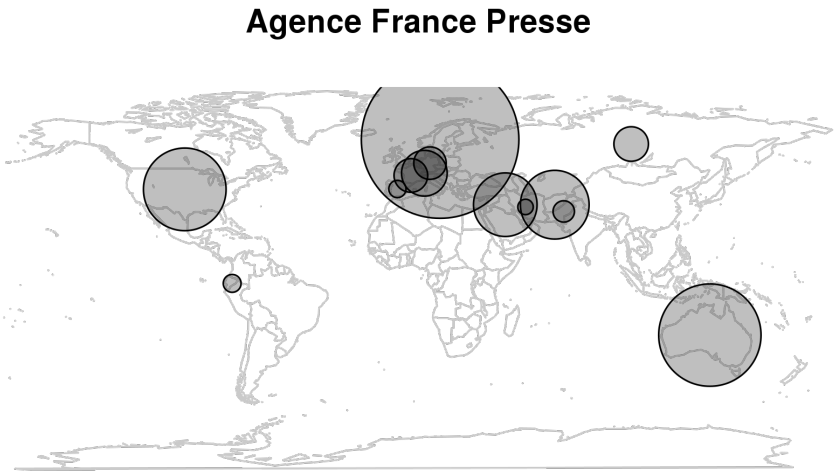
Figure 10 represents the resulting maps (the areas of circles are proportional to frequencies). It helps in figuring out the geographical dimension of the way news agencies framed the Assange case. Similarities between the three agencies are very striking. No alternative geographical framing occured during the two studied years outside the episodic frame focusing mainly on Sweden and the UK, secondarily on the US and Australia and only thirdly on other countries mostly connected to Julian Assange due to various Wikileaks operations (Iraq, Afghanistan, Switzerland, Russia, Germany, etc.). But when looked closely at those maps slightly differ. For instance, over-reporting of agencies' origin countries is observable, as well as some peculiarities such as the importance of references to Iran in Reuters' dispatches, Pakistan in AFP's ones, etc.[33]

32 The data frame used here is freely available for download at http://gothos.info/resources/. Many other data frames containing geographical data (city coordinates for instance) can be found easily on the Web.

33 The code is the following (this requires code from note 27 to be run first) :

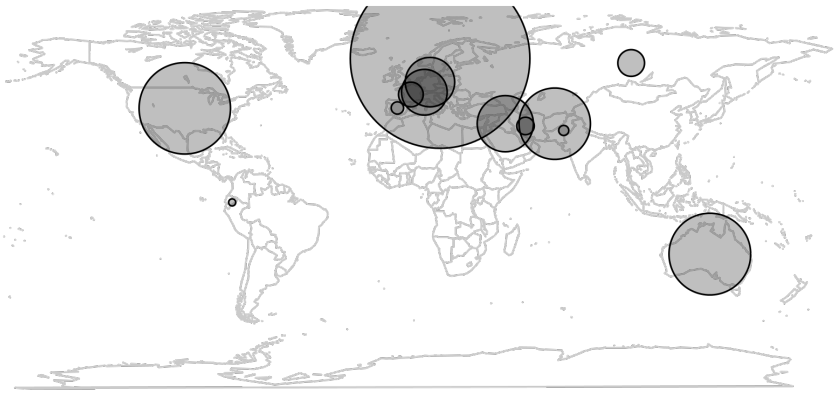
```
# Creation of a data frame containing all country frequencies
# (countries being designated with the two-letter ISO 3166 code that is used in the country centroids file)
freqs <- as.data.frame(freqs)
# Simplification of agency names
names(freqs)[1] <- "AFP"
names(freqs)[2] <- "AP"
names(freqs)[3] <- "Reuters"
# Importation of the country centroids file (previously downloaded on the hard disk)
centroids <- read.delim(file="country_centroids_primary.csv", header=TRUE)
dat <- merge(freqs, centroids, by.x="row.names", by.y="ISO3136")
# Drawing of the three maps
library(maps)
map(database="world", col="grey")
title("Agence France Presse")
points(dat$LONG, dat$LAT, pch=21, cex=10*sqrt(dat$AFP), bg="#00000040")
map(database="world", col="grey")
title("Associated Press")
```

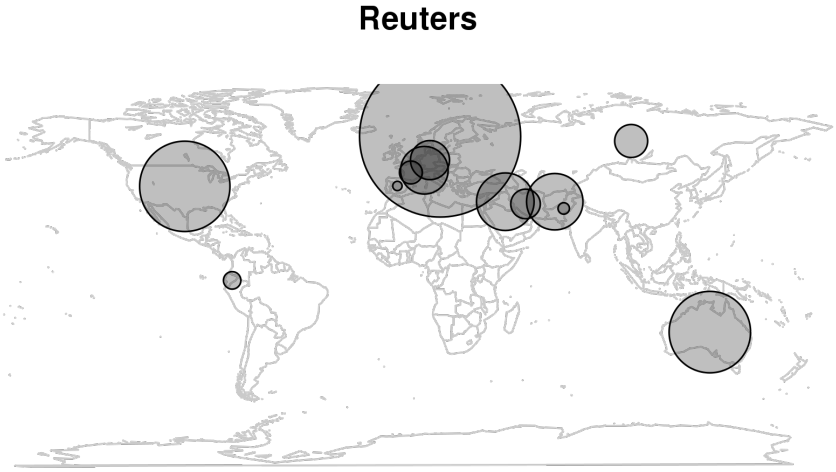
Fig. 10 — Mapping J. Assange in the three agencies




```
points(dat$LONG, dat$LAT, pch=21, cex=10*sqrt(dat$AP), bg="#00000040")
map(database="world", col="grey")
title("Reuters")
points(dat$LONG, dat$LAT, pch=21, cex= 10*sqrt(dat$Reuters), bg="#00000040")
```

Associated Press





Conclusion

Due to the peculiarities of news agencies dispatches that obey a very routinized production process and reach a global audience, the coverage of Julian Assange in 2010-2011 is very similar among the three

world agencies.[34] This of course is an interesting result if one looks at broad differences in the way that news is shaped at the global level for events with a clearly episodic nature. But it is also very interesting to notice that using sophisticated tools with fine tuning options can help to identify differences in the dispatches of the three agencies. These small differences tend to oppose AFP on one side and mostly AP on the other. The French agency tended to be less concerned with the legal procedure and the sexual case and more with the geopolitical and financial aspects of the whistleblower's situation. It also provided a more personal description of Assange, relying less on official reports and more on accounts by his relatives.

We hope to have provided with this corpus analysis evidence that opening the black box of proprietary text mining solutions is of major interest for media studies. With R.TeMiS we propose to explore two new dimensions in text mining. The automation of corpus construction and management procedures is first: easily gathering media content with relevant metadata in a reliable way is key to successful text mining in media studies. The extension of the range of statistical tools available through R coding is the second one. Providing the social scientist with standard statistical methods developed in R and facilitating the invention of new tools based on other R packages is also something that media studies — and every other field of the HSS — can benefit from.

34 It can be seen in the CA on the unaggregated DTM that the three agencies are located near the center of the first plane (Fig. 1) whereas monthly breaks on the Date variable have more scattered coordinates (not shown). This suggests that the corpus is structured by its time rather than source structure.

Appendix

Appendix 1 — 25 most specific terms for each source

Agence France Presse

	% Term/Level	% Level/Term	Global %	Level	Global	t value	Prob.
assault	0.2342	83.9806	0.1509	173	206	Inf	0.0000
whistleblow	0.3127	90.5882	0.1868	231	255	Inf	0.0000
organis	0.1855	79.6512	0.1260	137	172	6.9114	0.0000
websit	0.7770	65.3759	0.6432	574	878	6.7515	0.0000
borgstroem	0.0514	100.0000	0.0278	38	38	6.4088	0.0000
rape	0.6362	65.8263	0.5231	470	714	6.3260	0.0000
warrant	0.4914	67.3469	0.3949	363	539	6.2132	0.0000
claim	0.3763	69.3267	0.2938	278	401	6.1697	0.0000
australian	0.5848	65.8537	0.4806	432	656	6.0734	0.0000
dollar	0.0663	92.4528	0.0388	49	53	5.9654	0.0000
enrag	0.0569	95.4545	0.0322	42	44	5.9522	0.0000
arrest	0.7039	63.9606	0.5956	520	813	5.6611	0.0000
world	0.3492	68.2540	0.2769	258	378	5.5504	0.0000
misconduct	0.0027	2.5000	0.0586	2	80	-10.0703	0.0000
sex	0.1719	30.7506	0.3026	127	413	-9.5486	0.0000
organ	0.0555	23.1638	0.1297	41	177	-8.3283	0.0000
elmer	0.0054	6.8966	0.0425	4	58	-7.5578	0.0000
borgstrom	0.0000	0.0000	0.0293	0	40	-7.5125	0.0000
aug	0.0000	0.0000	0.0286	0	39	-7.4098	0.0000
volunt	0.0095	10.4478	0.0491	7	67	-7.4073	0.0000
investig	0.2098	37.6214	0.3018	155	412	-6.6806	0.0000
inc	0.0000	0.0000	0.0227	0	31	-6.5328	0.0000
assang	3.1784	49.5568	3.4710	2348	4738	-6.3881	0.0000
offens	0.0041	7.1429	0.0308	3	42	-6.3297	0.0000
crime	0.1936	39.0710	0.2681	143	366	-5.7282	0.0000

Associated Press Newswires

	% Term/Level	% Level/Term	Global %	Level	Global	t value	Prob.
organ	0.3253	53.1073	0.1297	94	177	Inf	0.0000
sex	0.5745	40.1937	0.3026	166	413	Inf	0.0000
offens	0.1177	80.9524	0.0308	34	42	8.1259	0.0000
spill	0.1073	73.8095	0.0308	31	42	7.1528	0.0000
borgstrom	0.1038	75.0000	0.0293	30	40	7.1301	0.0000

wasn	0.0761	84.6154	0.0190	22	26	6.7247	0.0000
stockholm	0.3080	41.5888	0.1568	89	214	6.6734	0.0000
complaint	0.0865	69.4444	0.0264	25	36	6.0599	0.0000
consensu	0.1177	57.6271	0.0432	34	59	5.9758	0.0000
defens	0.1661	47.5248	0.0740	48	101	5.7897	0.0000
inc	0.0761	70.9677	0.0227	22	31	5.7751	0.0000
mayawati	0.0415	100.0000	0.0088	12	12	5.6488	0.0000
women	0.4603	33.4171	0.2916	133	398	5.6160	0.0000
seek	0.1938	42.4242	0.0967	56	132	5.4135	0.0000
stem	0.0865	59.5238	0.0308	25	42	5.2540	0.0000
disclosur	0.1592	44.6602	0.0755	46	103	5.2407	0.0000
organis	0.0000	0.0000	0.1260	0	172	-8.7000	0.0000
websit	0.3219	10.5923	0.6432	93	878	-8.2627	0.0000
whistleblow	0.0381	4.3137	0.1868	11	255	-7.5871	0.0000
bank	0.0485	5.1282	0.2000	14	273	-7.3793	0.0000
assault	0.0277	3.8835	0.1509	8	206	-7.0169	0.0000
cabl	0.3807	12.1951	0.6608	110	902	-6.9955	0.0000
miss	0.0035	1.1236	0.0652	1	89	-5.5320	0.0000
arrest	0.3876	13.7761	0.5956	112	813	-5.3834	0.0000
defenc	0.0035	1.1905	0.0615	1	84	-5.3295	0.0000

Reuters News

	% Term/Level	% Level/Term	Global %	Level	Global	t value	Prob.
elmer	0.1453	84.4828	0.0425	49	58	Inf	0.0000
misconduct	0.1779	75.0000	0.0586	60	80	Inf	0.0000
volunt	0.1630	82.0896	0.0491	55	67	Inf	0.0000
word	0.1571	64.6341	0.0601	53	82	7.5241	0.0000
miss	0.1541	58.4270	0.0652	52	89	6.6564	0.0000
femal	0.0741	75.7576	0.0242	25	33	5.9894	0.0000
plead	0.0534	90.0000	0.0147	18	20	5.9533	0.0000
author	0.4773	37.0968	0.3179	161	434	5.6826	0.0000
tax	0.0771	70.2703	0.0271	26	37	5.6711	0.0000
add	0.0623	77.7778	0.0198	21	27	5.6035	0.0000
lulzsec	0.0385	100.0000	0.0095	13	13	5.5691	0.0000
disc	0.0474	88.8889	0.0132	16	18	5.5153	0.0000
data	0.1453	50.5155	0.0711	49	97	5.3693	0.0000
cutler	0.0326	100.0000	0.0081	11	11	5.0599	0.0000
factbox	0.0326	100.0000	0.0081	11	11	5.0599	0.0000
updat	0.0534	75.0000	0.0176	18	24	4.9778	0.0000
theatr	0.0356	92.3077	0.0095	12	13	4.8839	0.0000
bank	0.3083	38.0952	0.2000	104	273	4.8358	0.0000
rape	0.2490	11.7647	0.5231	84	714	-8.6343	0.0000
whistleblow	0.0385	5.0980	0.1868	13	255	-8.3088	0.0000
claim	0.1067	8.9776	0.2938	36	401	-8.0251	0.0000
victim	0.0030	1.2821	0.0571	1	78	-5.6887	0.0000
gillard	0.0000	0.0000	0.0425	0	58	-5.2640	0.0000
women	0.1660	14.0704	0.2916	56	398	-5.1589	0.0000

Appendix 2 — Chosen terms frequencies according to the origin of the dispatches

, , assault

	% Term/Level	% Level/Term	Global %	Level	Global	t value	Prob.
Agence France Presse	0.234	83.981	0.151	172.000	206.000	Inf	0.000
Associated Press Newswires	0.028	3.883	0.151	8.000	206.000	-7.017	0.000
Reuters News	0.074	12.136	0.151	25.000	206.000	-4.403	0.000

, , rape

	% Term/Level	% Level/Term	Global %	Level	Global	t value	Prob.
Agence France Presse	0.64	65.83	0.52	469	714	6.33	0.00
Associated Press Newswires	0.55	22.41	0.52	159	714	0.77	0.22
Reuters News	0.25	11.76	0.52	84	714	-8.63	0.00

, , molest

	% Term/Level	% Level/Term	Global %	Level	Global	t value	Prob.
Agence France Presse	0.2477	57.7287	0.2322	182	317	1.2360	0.1082
Associated Press Newswires	0.2942	26.8139	0.2322	84	317	2.3398	0.0096
Reuters News	0.1453	15.4574	0.2322	49	317	-3.9392	0.0000

, , offens

	% Term/Level	% Level/Term	Global %	Level	Global	t value	Prob.
Agence France Presse	0.0041	7.1429	0.0308	3	42	-6.3297	0.0000
Associated Press Newswires	0.1177	80.9524	0.0308	33	42	8.1259	0.0000
Reuters News	0.0148	11.9048	0.0308	5	42	-1.8310	0.0335

, , misconduct

	% Term/Level	% Level/Term	Global %	Level	Global	t value	Prob.
Agence France Presse	0.0027	2.5000	0.0586	2	80	-10.0703	0.0000
Associated Press Newswires	0.0623	22.5000	0.0586	17	80	0.1804	0.4284
Reuters News	0.1779	75.0000	0.0586	59	80	Inf	0.0000

, , crime

	% Term/Level	% Level/Term	Global %	Level	Global	t value	Prob.
Agence France Presse	0.1936	39.0710	0.2681	143	366	-5.7282	0.0000
Associated Press Newswires	0.3876	30.6011	0.2681	111	366	4.1783	0.0000

Appendix 3 — Geographical terms within the corpus according to the origin of the dispatches

, , Agence France Presse

	% Term/Level	% Level/Term	Global %	Level	Global	t value	Prob.
sweden	0.8839	52.4077	0.9128	653	1246	-1.1887	0.1173
swedish	0.6823	49.6552	0.7436	504	1015	-2.8284	0.0023
stockholm	0.1218	42.0561	0.1568	90	214	-3.4695	0.0003
britain	0.3763	55.0495	0.3700	277	505	0.3748	0.3539
british	0.3899	56.4706	0.3736	287	510	1.0235	0.1530
london	0.4968	50.5510	0.5319	367	726	-1.8949	0.0291
england	0.1327	59.3939	0.1209	97	165	1.2844	0.0995
english	0.0311	45.0980	0.0374	23	51	-1.1518	0.1247
uk	0.0203	40.5405	0.0271	15	37	-1.4912	0.0680
afghanistan	0.2396	62.3239	0.2081	176	284	2.7341	0.0031
afghan	0.0812	47.6190	0.0923	60	126	-1.3740	0.0847
iraq	0.1909	60.5150	0.1707	140	233	1.9018	0.0286
iraqi	0.0528	61.9048	0.0462	38	63	1.1153	0.1324
baghdad	0.0311	51.1111	0.0330	23	45	-0.2588	0.3979
australia	0.1272	61.4379	0.1121	93	153	1.7432	0.0406
australian	0.5848	65.8537	0.4806	431	656	6.0734	0.0000
america	0.0541	51.9481	0.0564	40	77	-0.2706	0.3933
american	0.1245	45.5446	0.1480	92	202	-2.3723	0.0088
washington	0.2139	55.6338	0.2081	157	284	0.4520	0.3257
york	0.0690	45.9459	0.0813	51	111	-1.6311	0.0514
usa	0.0027	10.0000	0.0147	2	20	-3.8939	0.0000
iceland	0.0596	63.7681	0.0505	43	69	1.4945	0.0675
russia	0.0393	47.5410	0.0447	29	61	-0.9031	0.1832
russian	0.0420	81.5789	0.0278	30	38	3.3620	0.0004
switzerland	0.0311	44.2308	0.0381	23	52	-1.2908	0.0984
swiss	0.1110	56.1644	0.1070	81	146	0.4112	0.3405
germani	0.0298	45.8333	0.0352	22	48	-1.0073	0.1569
german	0.0338	37.3134	0.0491	25	67	-2.6381	0.0042
berlin	0.0081	28.5714	0.0154	6	21	-2.1394	0.0162
franc	0.0365	61.3636	0.0322	26	44	0.8113	0.2086
french	0.0406	81.0811	0.0271	29	37	3.2441	0.0006
iran	0.0162	31.5789	0.0278	12	38	-2.6339	0.0042
pakistan	0.0311	82.1429	0.0205	22	28	2.8928	0.0019
ecuador	0.0217	66.6667	0.0176	15	24	1.0293	0.1517
spain	0.0203	75.0000	0.0147	14	20	1.6737	0.0471

, , Associated Press Newswires

	% Term/Level	% Level/Term	Global %	Level	Global	t value	Prob.
sweden	1.0590	24.5586	0.9128	305	1246	2.8629	0.0021
swedish	0.8237	23.4483	0.7436	237	1015	1.7322	0.0416
stockholm	0.3080	41.5888	0.1568	88	214	6.6734	0.0000
britain	0.3426	19.6040	0.3700	99	505	-0.8039	0.2107
british	0.2596	14.7059	0.3736	75	510	-3.6649	0.0001

london	0.5295	21.0744	0.5319	153	726	-0.0078	0.4969
england	0.1384	24.2424	0.1209	39	165	0.8765	0.1904
english	0.0311	17.6471	0.0374	9	51	-0.4164	0.3385
uk	0.0138	10.8108	0.0271	4	37	-1.3836	0.0832
afghanistan	0.2076	21.1268	0.2081	60	284	0.0695	0.5277
afghan	0.1384	31.7460	0.0923	39	126	2.6841	0.0036
iraq	0.1696	21.0300	0.1707	49	233	0.0440	0.5175
iraqi	0.0104	4.7619	0.0462	3	63	-3.4432	0.0003
baghdad	0.0346	22.2222	0.0330	9	45	0.0262	0.4895
australia	0.0934	17.6471	0.1121	27	153	-0.9670	0.1668
australian	0.3599	15.8537	0.4806	104	656	-3.3951	0.0003
america	0.0311	11.6883	0.0564	9	77	-1.9882	0.0234
american	0.2111	30.1980	0.1480	60	202	2.9439	0.0016
washington	0.2180	22.1831	0.2081	62	284	0.3586	0.3600
york	0.0900	23.4234	0.0813	25	111	0.4829	0.3146
usa	0.0173	25.0000	0.0147	4	20	0.1979	0.4216
iceland	0.0450	18.8406	0.0505	13	69	-0.2999	0.3821
russia	0.0346	16.3934	0.0447	10	61	-0.7429	0.2288
russian	0.0138	10.5263	0.0278	4	38	-1.4589	0.0723
switzerland	0.0346	19.2308	0.0381	10	52	-0.1398	0.4444
swiss	0.1073	21.2329	0.1070	30	146	-0.0627	0.5250
germani	0.0519	31.2500	0.0352	14	48	1.4976	0.0671
german	0.0727	31.3433	0.0491	20	67	1.8308	0.0336
berlin	0.0415	57.1429	0.0154	11	21	3.3972	0.0003
franc	0.0346	22.7273	0.0322	9	44	0.1042	0.4585
french	0.0069	5.4054	0.0271	2	37	-2.3708	0.0089
iran	0.0208	15.7895	0.0278	6	38	-0.5876	0.2784
pakistan	0.0069	7.1429	0.0205	2	28	-1.6868	0.0458
ecuador	0.0035	4.1667	0.0176	1	24	-1.9653	0.0247
spain	0.0104	15.0000	0.0147	3	20	-0.3531	0.3620

, , Reuters News

	% Term/Level	% Le-vel/Term	Global %	Level	Global	t value	Prob.
sweden	0.8508	23.0337	0.9128	287	1246	-1.3524	0.0881
swedish	0.8093	26.8966	0.7436	272	1015	1.5735	0.0578
stockholm	0.1038	16.3551	0.1568	35	214	-2.8650	0.0021
britain	0.3794	25.3465	0.3700	127	505	0.2868	0.3871
british	0.4358	28.8235	0.3736	146	510	2.0776	0.0189
london	0.6107	28.3747	0.5319	205	726	2.2231	0.0131
england	0.0800	16.3636	0.1209	27	165	-2.4858	0.0065
english	0.0563	37.2549	0.0374	18	51	1.8577	0.0316
uk	0.0534	48.6486	0.0271	17	37	2.9925	0.0014
afghanistan	0.1393	16.5493	0.2081	47	284	-3.2470	0.0006
afghan	0.0771	20.6349	0.0923	26	126	-0.9573	0.1692
iraq	0.1275	18.4549	0.1707	43	233	-2.1937	0.0141
iraqi	0.0623	33.3333	0.0462	20	63	1.4190	0.0780
baghdad	0.0356	26.6667	0.0330	11	45	0.1597	0.4366
australia	0.0949	20.9150	0.1121	32	153	-0.9962	0.1596
australian	0.3557	18.2927	0.4806	120	656	-3.8928	0.0000
america	0.0830	36.3636	0.0564	27	77	2.1677	0.0151
american	0.1453	24.2574	0.1480	49	202	-0.0549	0.4781
washington	0.1868	22.1831	0.2081	63	284	-0.9190	0.1791
york	0.1008	30.6306	0.0813	33	111	1.3235	0.0928
usa	0.0385	65.0000	0.0147	12	20	3.5957	0.0002
iceland	0.0356	17.3913	0.0505	12	69	-1.2886	0.0988
russia	0.0652	36.0656	0.0447	21	61	1.8551	0.0318
russian	0.0089	7.8947	0.0278	3	38	-2.4087	0.0080

switzerland	0.0563	36.5385	0.0381	18	52	1.7676	0.0386
swiss	0.0978	22.6027	0.1070	33	146	-0.4830	0.3145
germani	0.0326	22.9167	0.0352	11	48	-0.0928	0.4630
german	0.0623	31.3433	0.0491	20	67	1.1125	0.1330
berlin	0.0089	14.2857	0.0154	3	21	-0.8422	0.1998
franc	0.0208	15.9091	0.0322	7	44	-1.1945	0.1161
french	0.0148	13.5135	0.0271	5	37	-1.4290	0.0765
iran	0.0593	52.6316	0.0278	19	38	3.5393	0.0002
pakistan	0.0089	10.7143	0.0205	3	28	-1.5640	0.0589
ecuador	0.0208	29.1667	0.0176	6	24	0.3059	0.3799
spain	0.0059	10.0000	0.0147	2	20	-1.3050	0.0960

References

Abbott A (2000) Reflections on the Future of Sociology. Contemporary Sociology 29: 296-300.

Baker P (2006) Using corpora in discourse analysis: Continuum.

Becker R, Wilks A, Brownrigg R, et al. (2013) Maps: draw geographical maps. R package version 2.3-2.

Berelson B (1952) Content Analysis as a Tool of Communications Research, Glencoe (Ill.): Free Press.

Bolden R and Moscarola J (2000) Bridging the Quantitative-Qualitative Divide: The Lexical Approach to Textual Data Analysis. Social Science Computer Review 18: 450-460.

Bouchet-Valat M (2013) SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library. R package version 0.5. http://CRAN.R-project.org/package=SnowballC.

Bouchet-Valat M and Bastin G (2013) RcmdrPlugin.temis: Graphical Integrated Text Mining Solution. R package version 0.6.1. http://CRAN.R-project.org/package=RcmdrPlugin.temis.

Bouchet-Valat M and Bastin G (2013) RcmdrPlugin.temis, a Graphical Integrated Text Mining Solution in R. The R Journal 5(1): 188-197.

de Bonville J (2000) L'analyse de contenu des médias. De la problématique au traitement statistique, Bruxelles: De Boeck.

De Vreese CH (2005) News framing: Theory and typology. Information design journal+ document design 13: 51-62.

Demazière D (2005) Des logiciels d'analyse textuelle au service de l'imagination sociologique. Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique 85: 5-9.

Demazière D and Brossaud C (2006) Méthodes logicielles et réflexivité du sociologue. In: Brossaud C, Trabal P and van Meter K (eds) Analyses textuelles en sociologie. Logiciels, méthodes, usages. Rennes: Presses Universitaires de Rennes, 11-22.

Entman RM (1993) Framing: Toward Clarification of a Fractured Paradigm. Journal of Communication 43: 51-58.

Feinerer I (2008) An introduction to text mining in R. R News 8: 19-22.

Feinerer I (2011) Introduction to the tm Package Text Mining in R.

Feinerer I, Hornik K and Meyer D (2008) Text mining infrastructure in R. Journal of Statistical Software 25: 1-54.

Fox J (2005) The R Commander : A Basic-Statistics Graphical User Interface to R. Journal of Statistical Software 14.

Gamson WA and Modigliani A (1987) The changing culture of affirmative action. In: Braumgart R (ed) Research in Political Sociology. JAI Press, 137-177.

Gerbner G (1958) On Content Analysis and Critical Research in Mass Communication. Audio Visual Communication Review 6: 85-108.

Guerreau A (1989) Pourquoi (et comment) l'historien doit-il compter les mots? Histoire & mesure 4: 81-105.

Hall S (1978) Policing the Crisis. Mugging the State, and Law and Order, London: Macmillan Education.

Hey T and Trefethen A (2003) The Data Deluge: An e-Science Perspective. In: Berman F, Fox G and Hey T (eds) Grid Computing: Making the Global Infrastructure a Reality. Wiley & Sons, 809-824.

Iyengar S (1994) Is Anyone Responsible? How Television Frames Political Issues, Chicago: University of Chicago Press.

Krippendorff K (2004a) Content analysis: An introduction to its methodology: Sage Publications, Inc.

Krippendorff K (2004b) Reliability in Content Analysis. Human Communication Research 30: 411-433.

Krippendorff K and Bock MA (2008) The content analysis reader: SAGE Publications, Incorporated.

Lebart L and Salem A (1994) Statistique textuelle, Paris: Dunod.

Lebart L, Salem A and Berry L (1998) Exploring textual data, Dordrecht: Kluwer Academic.

Mills CW (1959) The sociological imagination, New York: Oxford University Press.

Muller C (1969) La statistique lexicale. Langue française: 30-43.

R Core Team (2013) R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Sarkar D (2008) Lattice: multivariate data visualization with R: Springerverlag New York.

Schafraad P, Wester F and Scheepers P (2006) Using ‘new’data sources for ‘old’newspaper research: developing guidelines for data collection. Communications 31: 455-467.

Sinclair J (2004) Trust the text: Language, corpus and discourse: Routledge.

Tognini-Bonelli E (2001) Corpus linguistics at work: John Benjamins Publishing Company.

Tuchman G (1972) Objectivity as strategic ritual. An examination of newsmens' notion of objectivity. American journal of sociology 77: 660-679.

Zeileis A and Grothendieck G (2005) zoo: S3 infrastructure for regular and irregular time series. arXiv preprint math/0505527.